The On-chip 3MB Subarray Based 3rd Level Cache on an Itanium™ Microprocessor

Don Weiss

John J. Wuu

Victor Chin[1]

Hewlett-Packard Company, Fort Collins, CO

[1]Intel Corporation, Santa Clara, CA

Outline

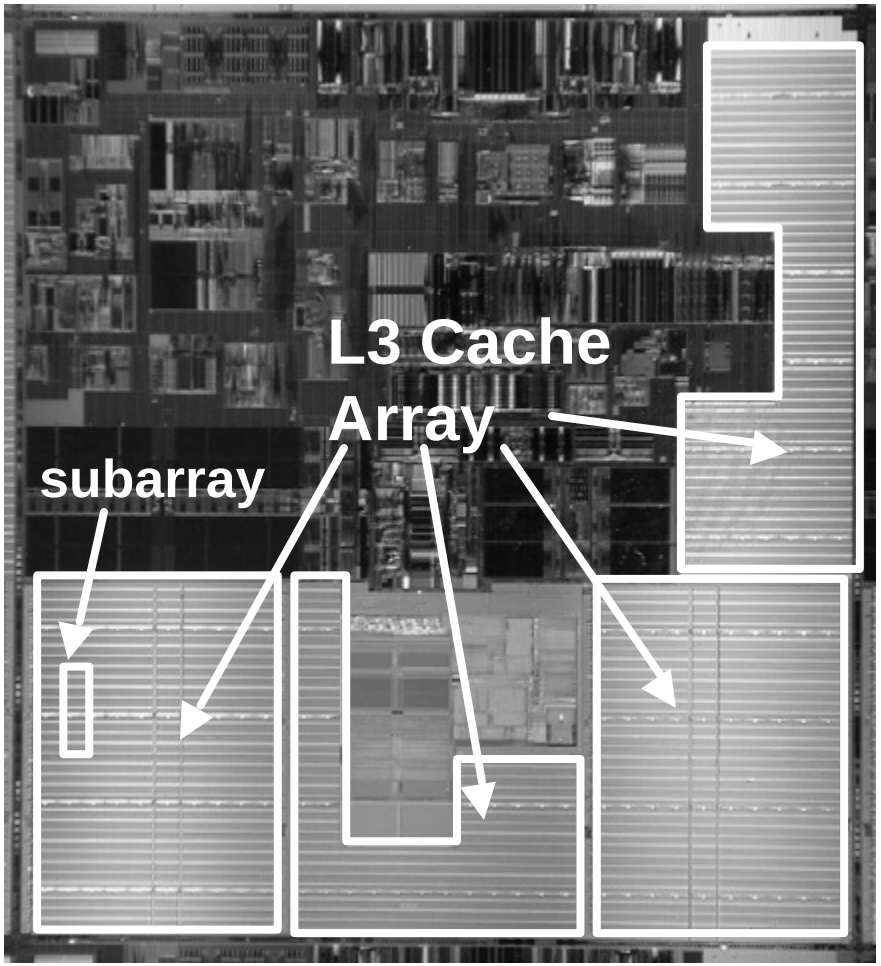
- Background
- Subarray Design Style
- Array Density
- Read Circuits
- Write Circuits
- Other Features
- Summary

Background

- 3 MB
- 12-way, single ported
- 1024b line size
- 3rd level cache on McKinley
  - 1st level cache optimized for latency
  - 2nd level cache optimized for bandwidth
  - 3rd level cache optimized for size
- Must efficiently utilize available area

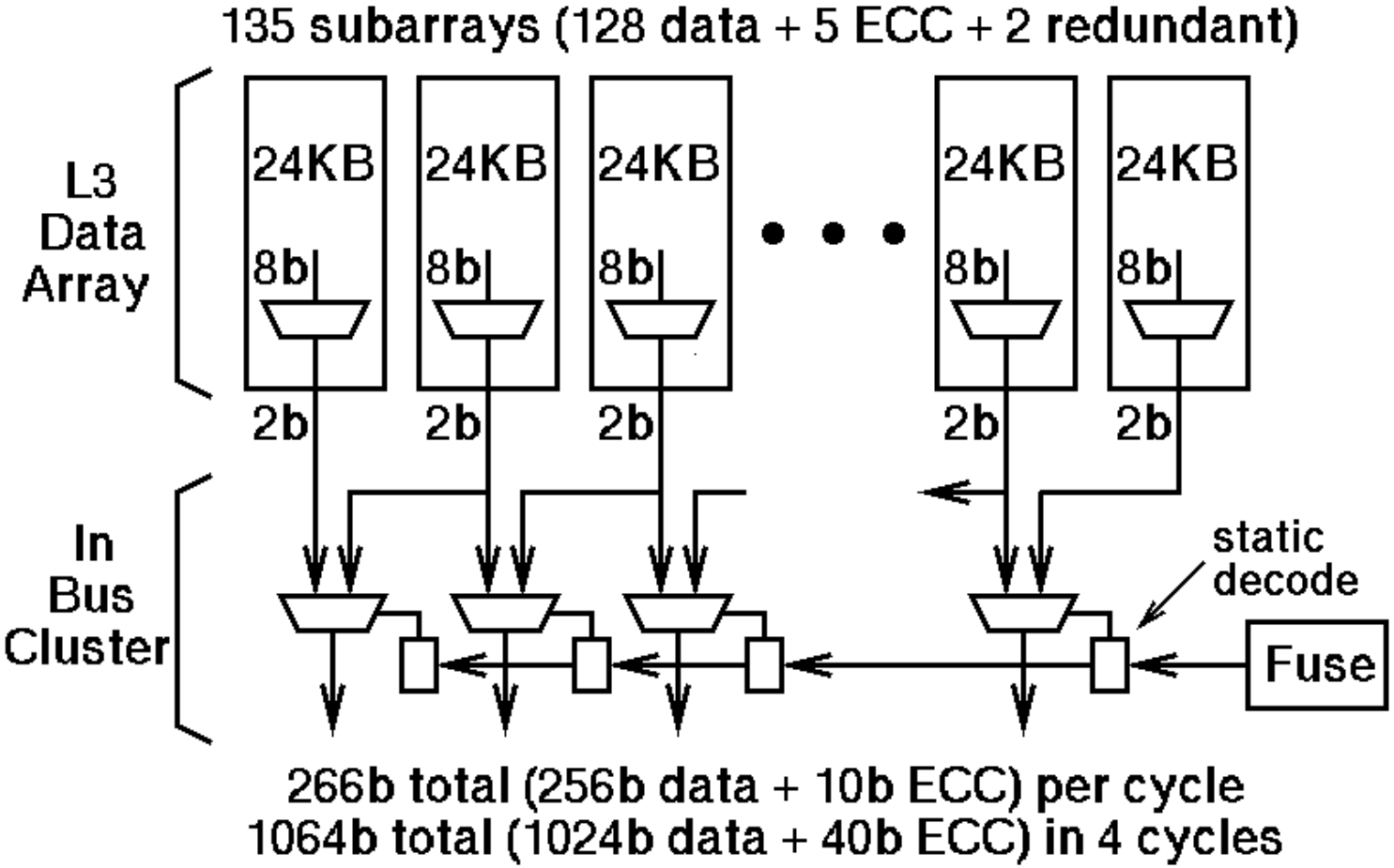
Subarray Design Style

- 3 MB cache broken down to 135 24KB subarrays
  - 128 data subarrays
  - 5 ECC subarrays
  - 2 redundant subarrays
- Efficiently fills up available space (175mm^2)



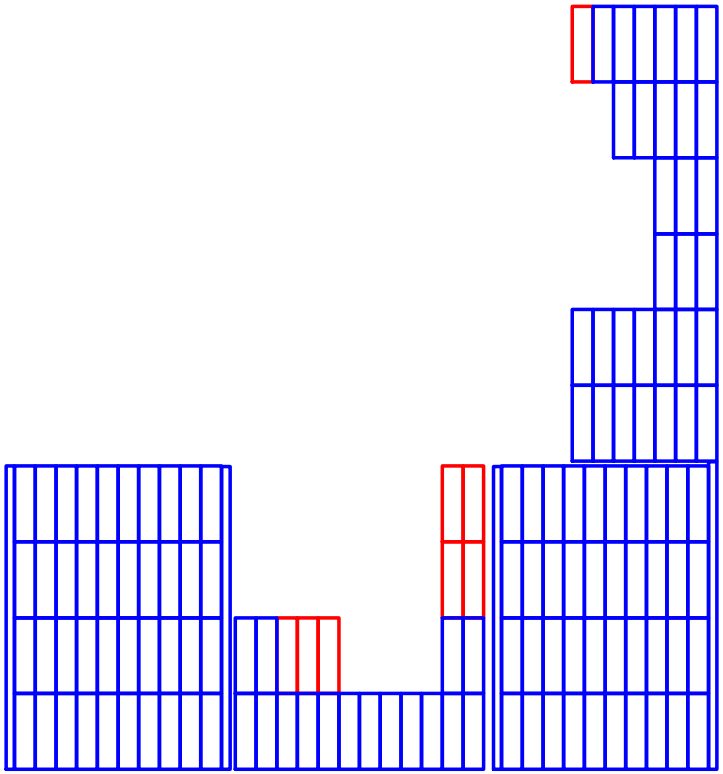
Subarray Design Style

Subarray arrangement

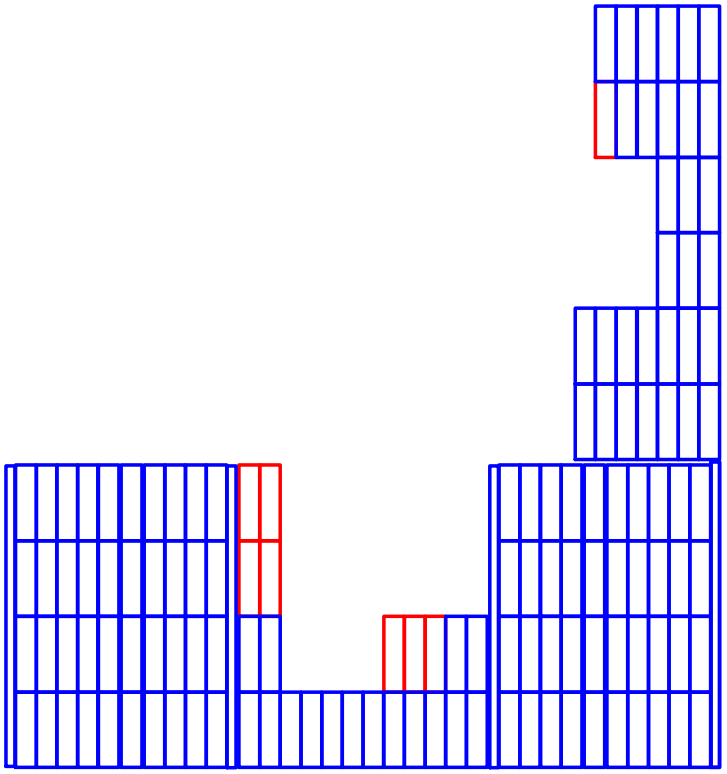


Subarray Design Style

Flexibly adapts to floor plan changes



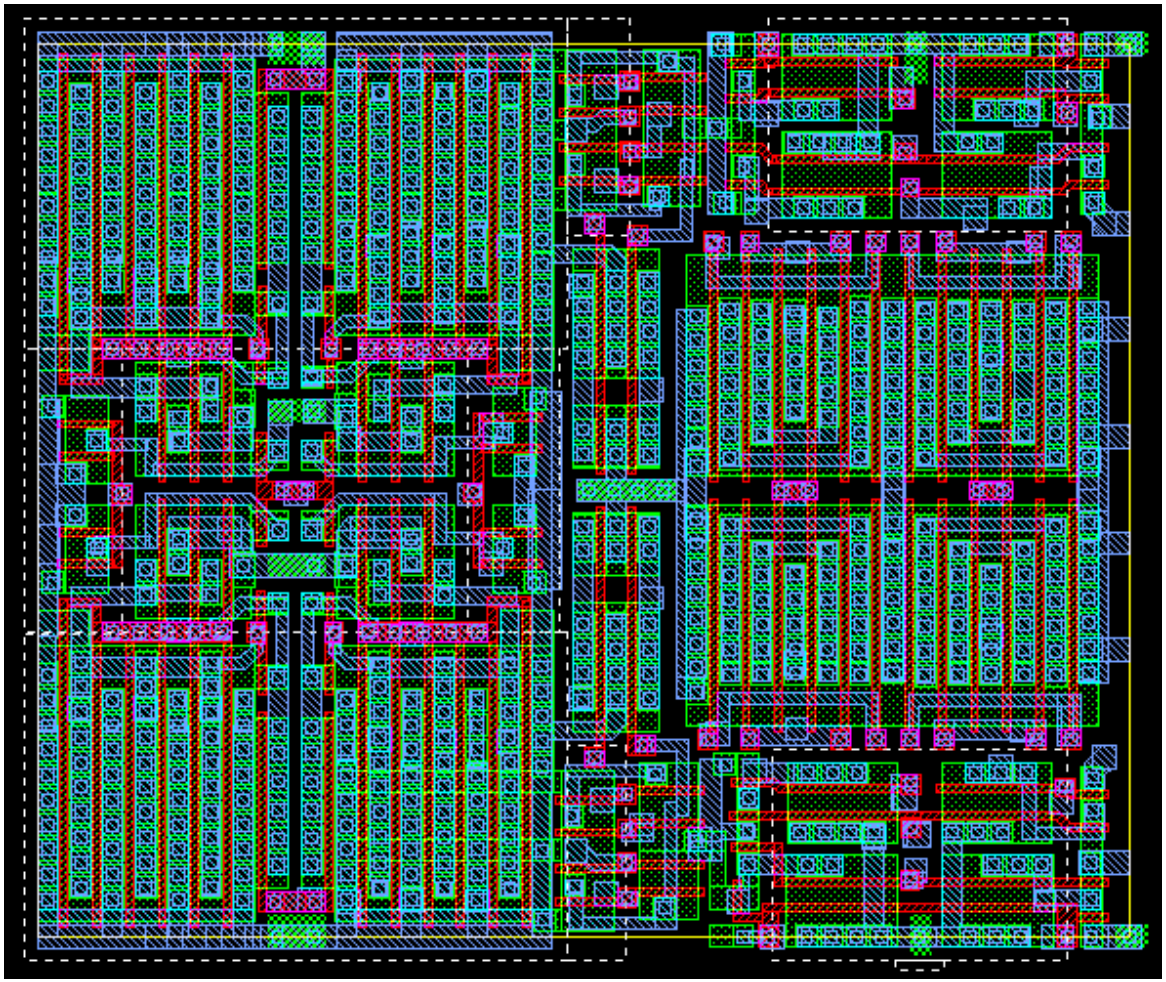
Floor plan snap shot



Final floor plan

Array Density

Fully custom cells and layout

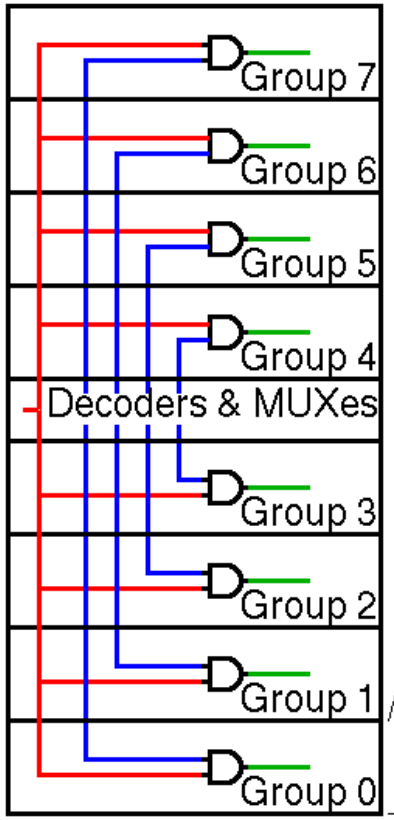


- Circuit and layout designers worked closely to produce compact layouts

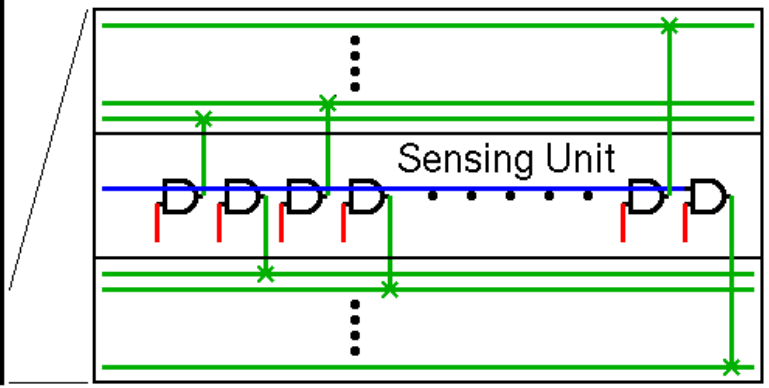
Array Density

- Load to driver size is non-linear as wires get longer
- Partition wires with distributed decoding to hit RC “sweet spot”
- Small decoders and drivers occupy only 3% of subarray area
- To reduce need for repeater channels, 2.5% of subarray area is allocated for global repeaters
- Achieved 85% array density for subarrays

Array Density
Distributed decoder scheme



- 8 groups per subarray
- 96 word lines, 8 column selects per group
- 1st level decoders in middle of subarray
- 2nd level decoders hidden in sensing units



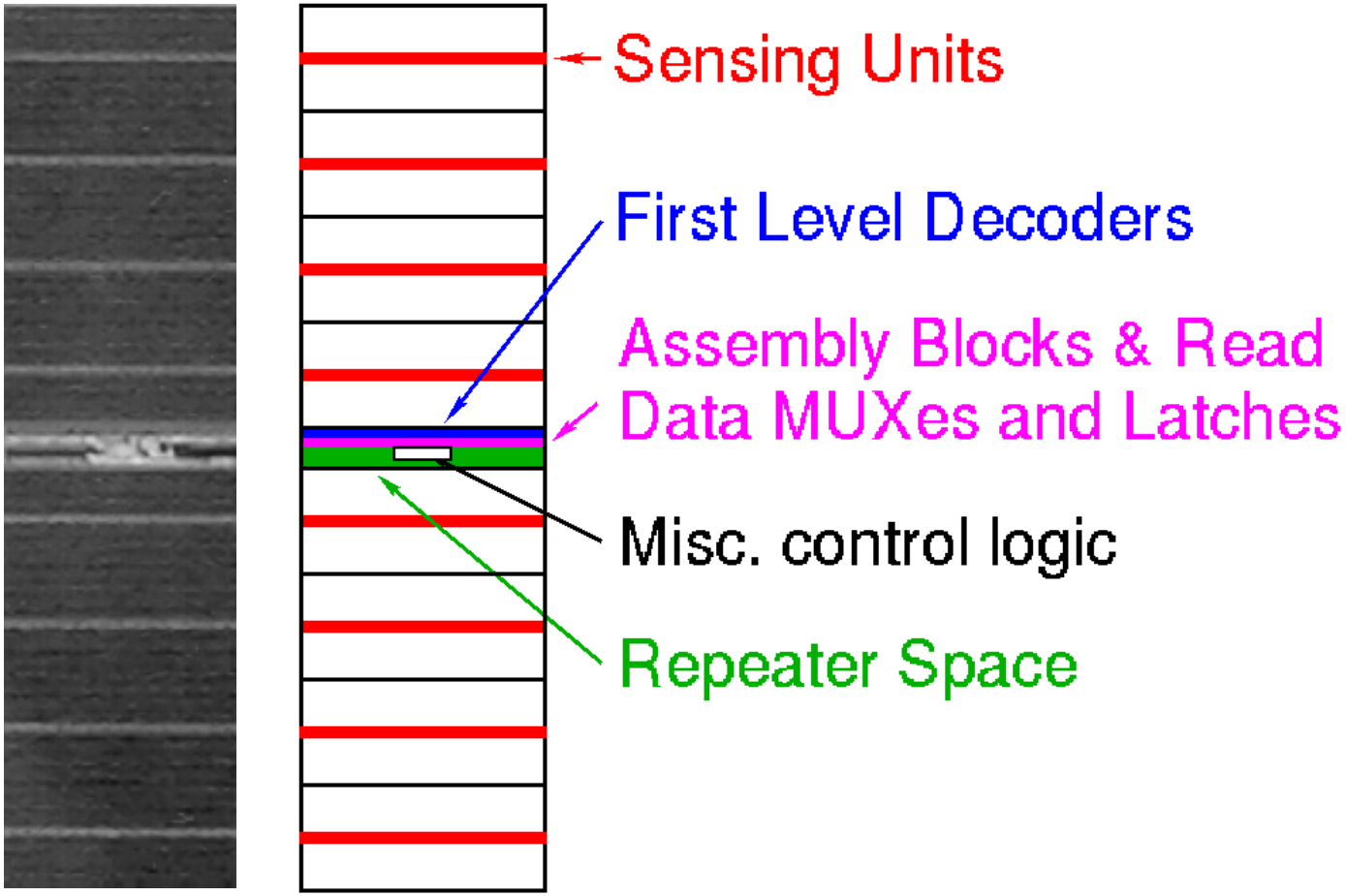
Group clock

Global word lines

Local word lines

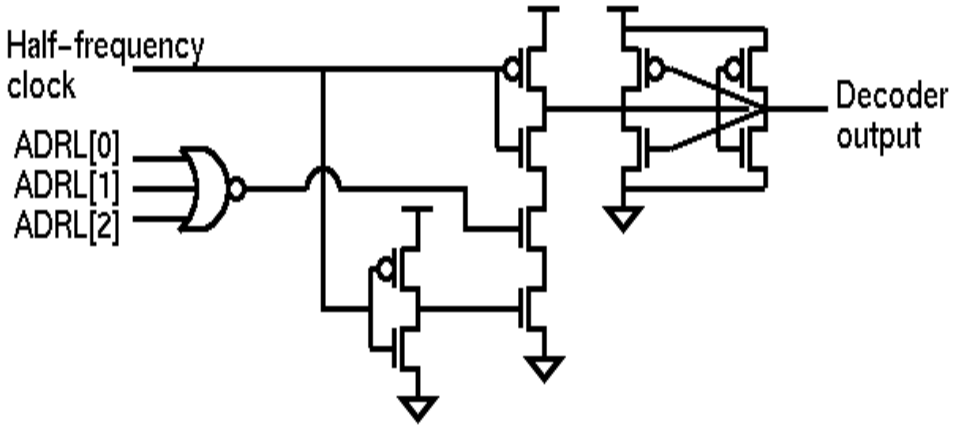
Array Density

Subarray layout

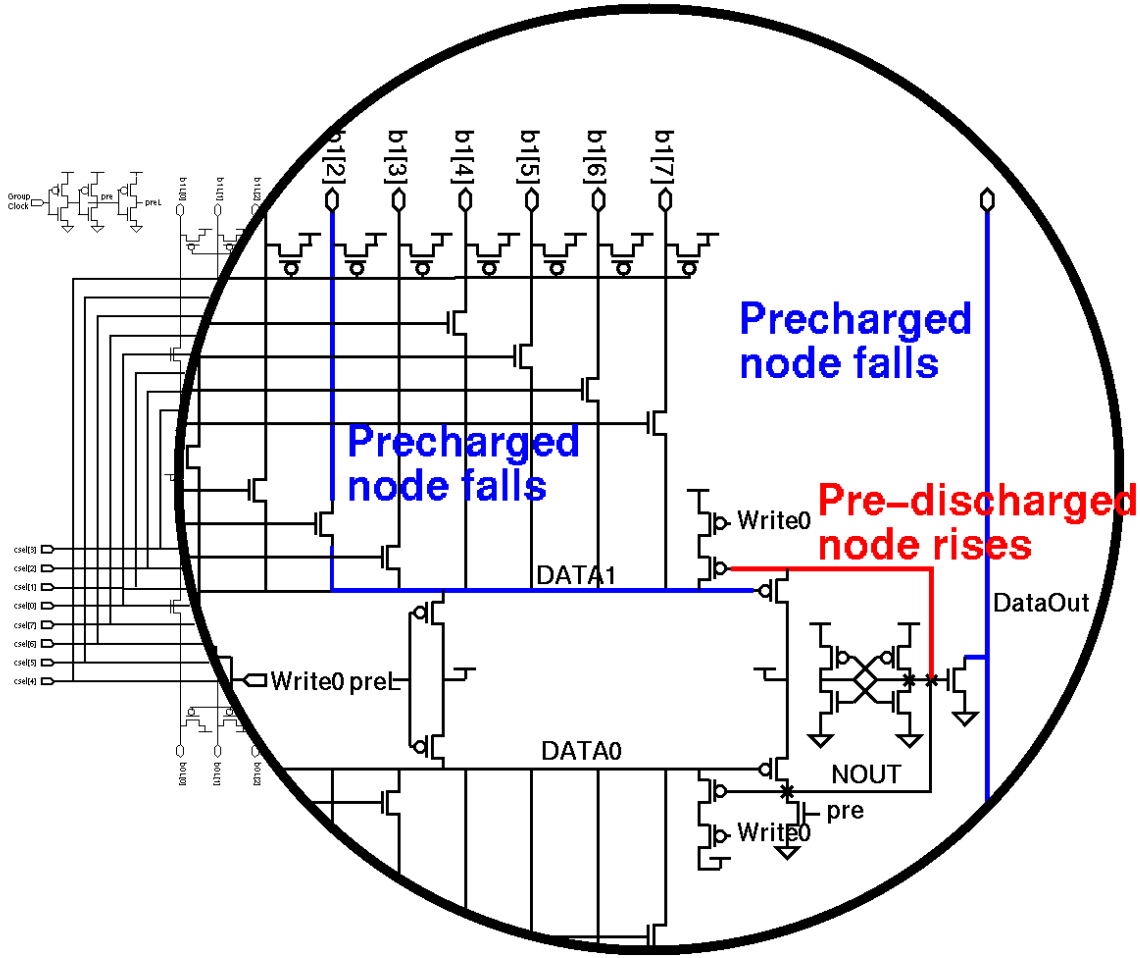


Read Circuits

- 1st level decoder: decodes and latches
  - 8 group clocks
  - 96 global word lines
  - 8 global column selects
- 2nd level decoders: static AND gates for
  - Local word lines
  - Local column selects



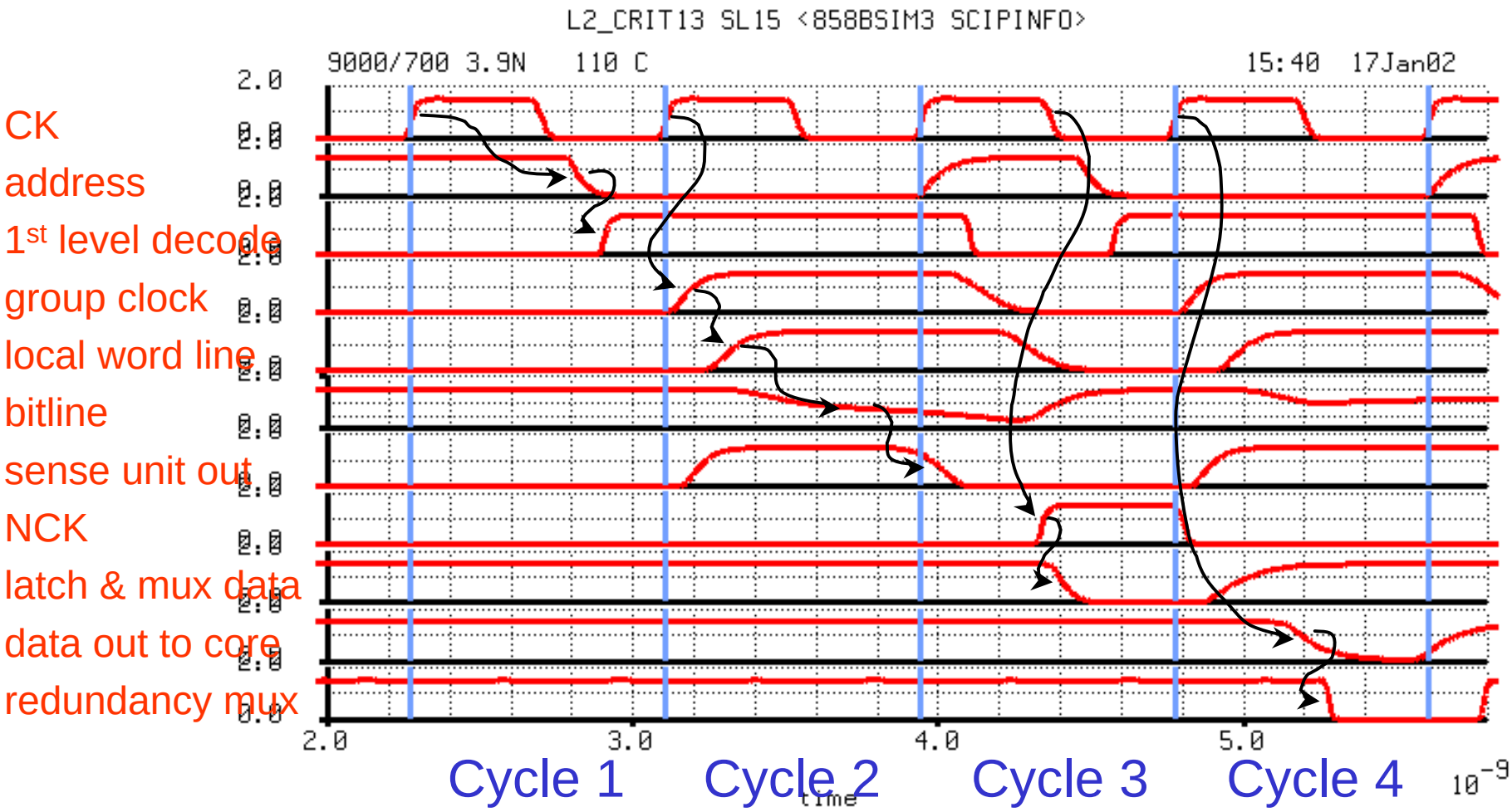
Read Circuits



- Single ended sensing
- Sensing unit output latched in middle of subarray
- Data piped to CPU in 4 cycles, critical chunk first

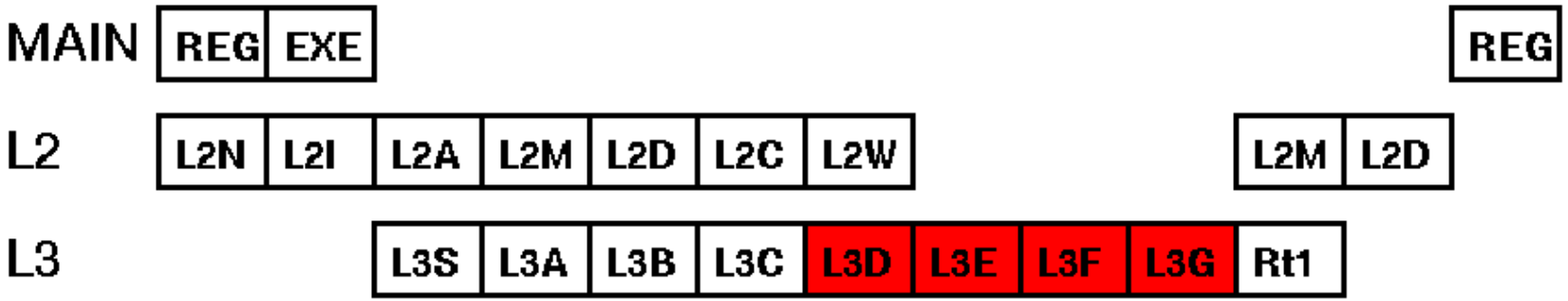
Read Circuits

Timing diagram



Read Circuits

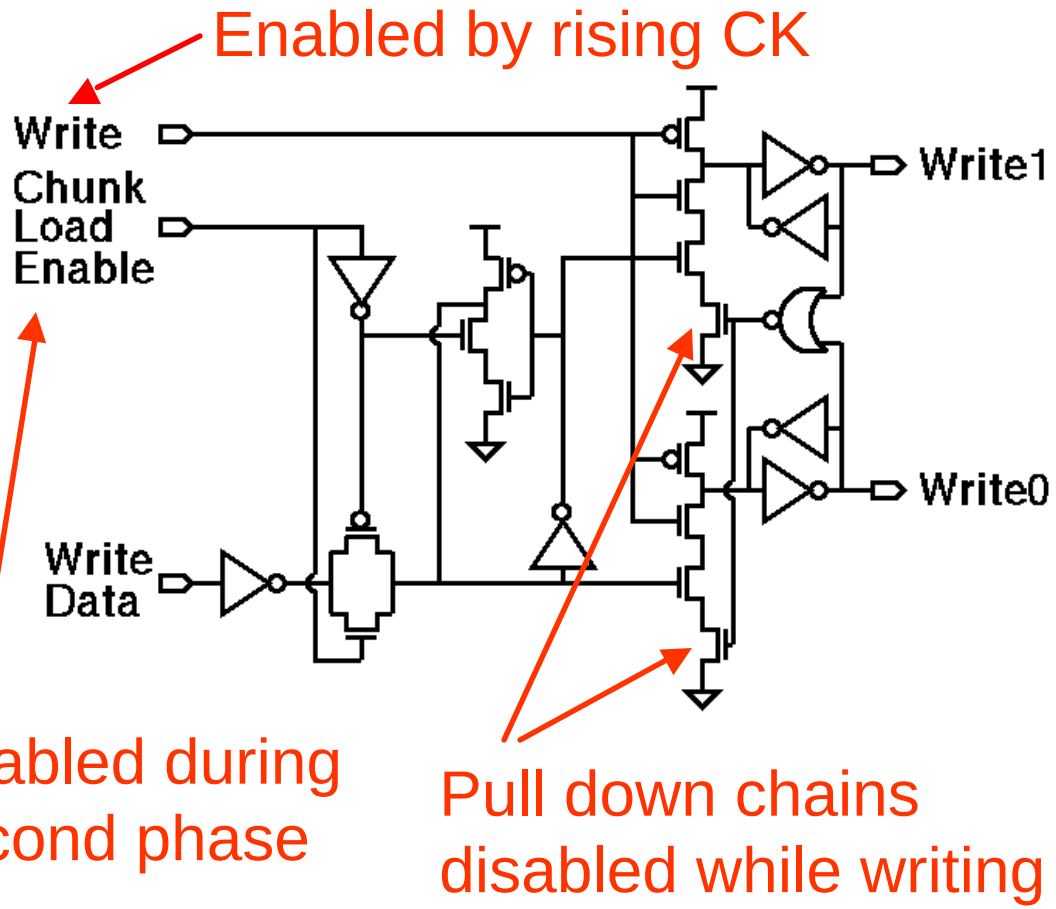
Pipeline diagram



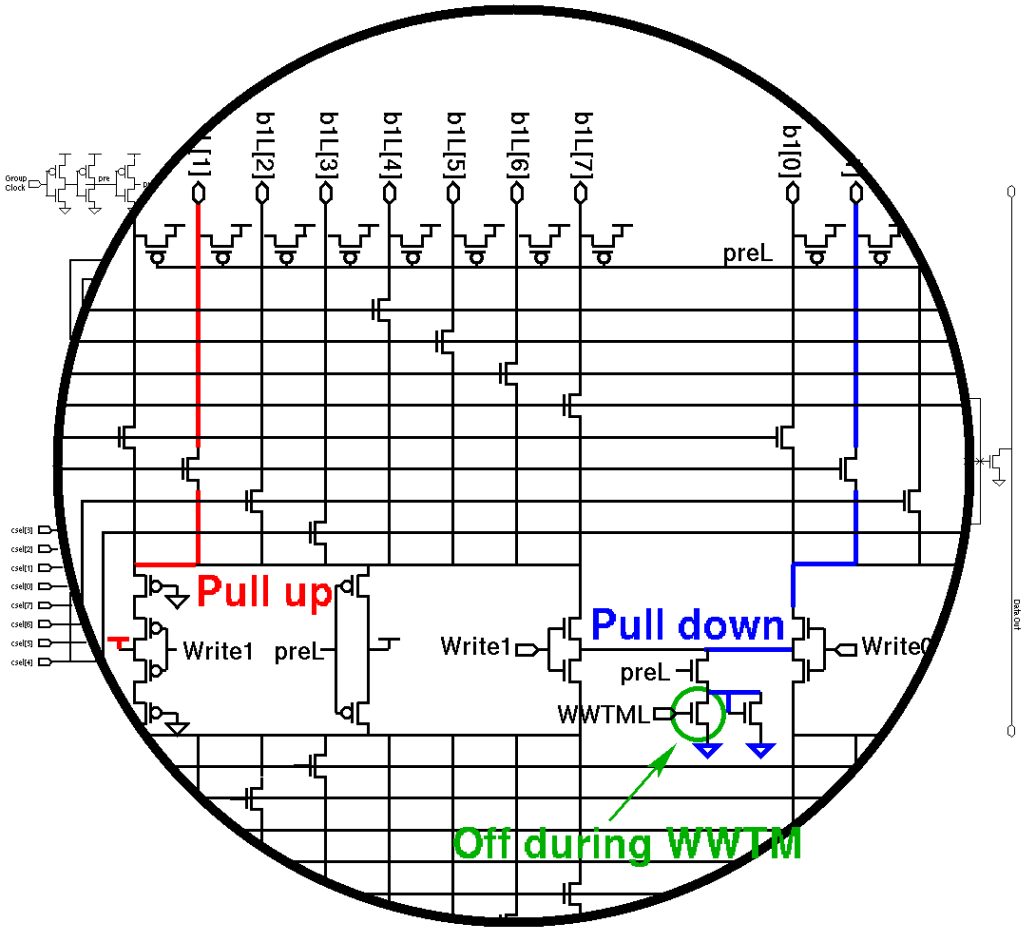
L3S	Request issued from L2 to L3	L3D-L3G	L3 data array access
L3A-L3C	L3 tag access and hit compare	Rt1	Critical chunk return to L2 from BC

Write Circuits

- Write data latched by assembly block in 4 cycles
- Feedback circuit allows latching new data while writing



Write Circuits



- Dual-ended write into SRAM cell from sensing unit
- Weak Write Test Mode (WWTM) incorporated into write circuit

Other Features

- ECC:
  - Single bit correct
  - Double bit detect
- Redundancy: 2 redundant subarrays capable of repairing defects ranging from single cell to full subarray failures
- DAT: Can be characterized and screened through Direct Access Test
- BIST: Burn-in toggle coverage and power-on self test
- All control signals scannable through JTAG

Summary

- 3 MB Level 3 Cache
- Flexible subarray design style
- Achieved 85% array efficiency
- Can perform 1 read and 1 write every 4 cycles
- 64 GB/s total bandwidth (at 1.0 GHz)
- Characterized to operate above 1.2 GHz at 1.5V, 110°C